Published in final edited form as:
*Nature*. 2008 October 2; 455(7213): 661–664. doi:10.1038/nature07390.

# Direct Evidence of Extensive Diversity of HIV-1 in Kinshasa by 1960

**Michael Worobey**[1,*], **Marlea Gemmel**[1], **Dirk E. Teuwen**[2,3], **Tamara Haselkorn**[1], **Kevin Kunstman**[4], **Michael Bunce**[5], **Jean-Jacques Muyembe**[6,7], **Jean-Marie M. Kabongo**[6], **Raphaël M. Kalengayi**[6], **Eric Van Marck**[8], **M. Thomas P. Gilbert**[1,9], and **Steven M. Wolinsky**[2]

[1]Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721

[2]Division of Infectious Disease, sanofi pasteur, F-69367 Lyon Cedex 07, France

[3]Rega Institute, Leuven, Belgium

[4]The Feinberg School of Medicine, Northwestern University, Chicago, IL 60611

[5]Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, Perth, Western Australia, Australia

[6]University of Kinshasa, Kinshasa, Democratic Republic of the Congo

[7]National Institute for Biomedical Research, National Laboratory of Public Health, Kinshasa B.P. 1197, Democratic Republic of the Congo

[8] Department of Pathology, University Hospital, University of Antwerp, Antwerp, Belgium

## Abstract

Human immunodeficiency virus type 1 (HIV-1) sequences that pre-date the recognition of AIDS are critical to defining the time of origin and the timescale of virus evolution[1,2]. A viral sequence from 1959 (ZR59) is the oldest known HIV-1 infection[1]. Other historically documented sequences, important calibration points to convert evolutionary distance into time, are lacking, however; ZR59 is the only one sampled prior to 1976. Here we report the amplification and characterization of viral sequences from a Bouin's-fixed paraffin-embedded lymph node biopsy specimen obtained in 1960 from an adult female in Léopoldville, Belgian Congo (now Kinshasa, Democratic Republic of the Congo [DRC]), and we use it to conduct the first comparative evolutionary genetic study of early pre-AIDS epidemic HIV-1 group M viruses. Phylogenetic analyses position this viral sequence (DRC60) closest to the ancestral node of subtype A (excluding A2). Relaxed molecular clock analyses incorporating DRC60 and ZR59 date the M group's most recent common ancestor near the beginning of the 20$^{th}$ century. The sizeable genetic distance between DRC60 and ZR59 directly demonstrates that diversification of HIV-1 in west-Central Africa occurred long before the recognized AIDS pandemic. The recovery of viral gene

[*]To whom correspondence should be addressed. worobey@email.arizona.edu..
[9]Current address: Centre for Ancient Genetics, Niels Bohr Institute and Biological Institute, University of Copenhagen, Copenhagen DK-2100, Denmark.

**Author Contributions** M.W., D.E.T., S.M.W., and M.T.P.G designed the study. M.G., T.H., K.K. and M.T.P.G. performed digestion/extraction, PCR, qPCR, cloning, and sequencing experiments. M.G. and M.B. designed primers. D.E.T., J-J.M., E.V.M., J-M.M.K and R.M.K. organized and provided samples. M.W. analyzed the data, performed the phylogenetic analyses, and wrote the paper. S.M.W. contributed to the analyses and writing. All authors discussed the results and commented on the manuscript.

The authors declare no competing financial interests.

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

The sequences reported in this study have been deposited in GenBank under accession numbers EU580739-EU580854 and EU589211-EU589236.

sequences from decades-old paraffin-embedded tissues opens the door to a detailed paleovirological investigation of the evolutionary history of HIV-1 that is not accessible by other methods.

---

We screened twenty-seven tissue blocks (8 lymph node, 9 liver, and 10 placenta) obtained from Kinshasa between 1958 and 1960 by RT-PCR; one lymph node biopsy specimen contained HIV-1 RNA. Viral nucleic acids were extracted from this specimen using protocols optimized for the recovery of nucleic acids from ancient or degraded samples[3,4]. After reverse transcription, 12 out of the 14 short HIV-1 cDNA fragments attempted (**Fig. 1A**) were amplified by PCR using a panel of conserved primer pairs from different regions of the viral genome (**Table S1**). Each PCR-product DNA was cloned and sequenced. Sequences were reproducible after repeated extractions and not the result of PCR contamination (see **Fig. 1A** and **Table S1** for fragment designations). The results were confirmed independently in two laboratories (**Figs. 1B** and **S1**), with the second laboratory successfully identifying the positive 1960 specimen in a blinded assay. The short fragments of the 1960 sample were found to be subtype A and not to be a mosaic of contemporary sequences (see supplementary information for a detailed discussion of the authenticity of the 1960 sequences). Consensus nucleotide sequences from these short HIV-1 fragments were concatenated for study. The analyses included reference sequences from the Los Alamos National Laboratory HIV sequence database and sequences recovered as part of this study from three paraffin-embedded tissue specimens collected from AIDS patients in Belgium and Canada between 1981 and 1997.

HIV-1 sequences were analyzed in MrBayes v3.1.2[5] using an unconstrained (no molecular clock enforced) Bayesian Markov chain Monte Carlo (BMCMC) method. The phylogenetic analyses confirmed that the DRC60 consensus sequences from the two laboratories were derived from a single patient (uncorrected pairwise distance 1.4%). The sequences were positioned close to the ancestral node of the subtype A lineage (excluding sub-subtype A2), forming a monophyletic clade with three modern sequences from the DRC (**Figs. 1B** and **S1**). Assuming a similar rate of evolution along all branches on a tree, the divergence between two sequences reflects the time elapsed since their shared ancestor. As predicted, the DRC60 sequences had a shorter branch length to the A/A1 ancestral node than the contemporary subtype A viruses sampled from the same geographic region ($P = 1.0$).

We validated the time of origin of the 1960 sequence by comparisons of the predicted date to the documented date. With the DRC60 date treated as an unknown, we calculated an evolutionary rate based on the distribution of branch lengths on the unconstrained phylogenetic trees sampled by MrBayes. To limit the effects of evolutionary rate differences between clades and uncertainties in rooting the HIV-1 M group phylogeny, we focused on the subtype A/A1 subtree (**Fig. S1**) and analyzed root-to-tip branch lengths relative to the sampling year. The mean estimates for the year of origin of the DRC60 consensus sequences from the University of Arizona and Northwestern University laboratories were 1959 (95% highest probability distribution 1902-84) and 1959 (95% HPD 1915-85), respectively, corroborating the authenticity of the DRC60 sequences and the existence of clock-like signal in our data set (see below). Despite initial indications that recombination might seriously confound phylogenetic dating estimates[6], subsequent work has suggested that recombination is not likely to systematically bias HIV-1 dates in one direction or the other, though it is expected to increase variance[7]. The close match between the predicted and actual dates of both ZR59[2] and DRC60 provides support for this view and an unambiguous indication that HIV-1 evolves in a fairly reliably clock-like fashion.

The uncorrected pairwise distance between DCR60 and ZR59 in their overlapping *env* region was 11.7% (**Fig. 1C**). This genetic distance is greater than 99.2% of within-subtype comparisons (within-subtype difference range 0.01-0.15; between-subtype difference range 0.05-0.18). Since each subtype represents several decades of independent evolution in the human population[2,8], the extensive divergence between DRC60 and ZR59 indicates that the HIV-1 M group founder virus began to diversify in the human population (and that HIV-1 likely entered Kinshasa) decades before 1960.

We applied a relaxed clock BMCMC coalescent framework as implemented in BEAST v1.4.7[9] to estimate the time to the most recent common ancestor (TMRCA) of the HIV-1 M group. This approach robustly incorporates phylogenetic uncertainty and accounts for the possibility of variable substitution rates among lineages and differences in the demographic history of the virus, sampling phylogenies and parameter estimates in proportion to their posterior probability[10]. As with other studies of HIV-1[11], comparisons of the marginal likelihoods of strict versus relaxed clock models (both of which are implemented in BEAST) indicated overwhelming support for relaxed clocks (data available upon request). Hence, the use of strict clock models with these data would be inappropriate and would likely yield misleadingly small error estimates with regard to both timing and substitution rates.

Using substitution rates calibrated with sequences sampled at different time points, we obtained a posterior distribution of rooted tree topologies with branch lengths in unit time (**Figs. 2** and **S2**). The median estimated substitution rate for the concatenated subregions of the *gag-pol-env* genes was $2.47 \times 10^{-3}$ substitutions/site/year (95% HPD $1.90\text{-}2.95 \times 10^{-3}$). The inclusion of the 1959 and 1960 sequences appeared to improve estimation of the TMRCA of the M group (**Table 1**), limiting the influence of the coalescent tree prior on the posterior TMRCA distributions compared with the data set that excluded these earliest cases of HIV-1. With DRC60 and ZR59 included, the different demographic/coalescent models gave highly consistent results, with tighter and more similar date ranges compared to the analyses that excluded them and 95% HPDs that extend no later than 1933. The best-fit model incorporated a constant population size demographic model (TMRCA 1921, 95% HPD 1908-1933). The model with a general, non-parametric prior (the Bayesian skyline plot tree prior)[12,13] that indicated a more complex (and biologically plausible) demographic history (**Fig. S3**) had a statistically indistinguishable degree of support (TMRCA 1908, 95% HPD 1884-1924). Moreover, the population expansion demographic model[9], which was a slightly worse fit to the data compared to the constant population and Bayesian skyline plot models, could not be rejected given the Bayes factors comparison of models (**Table 1**). The inability to strongly reject the model with a constant population size prior is counterintuitive since it is clear that the HIV-1 population size has increased dramatically. We speculate that this finding might be due to the simplest model providing a good fit to a relatively short, information-poor alignment, in comparison with more parameterized models.

Acid-containing fixatives such as Bouin's solution can cause base modifications of nucleic acids, leading to the generation of erroneous bases in sequences derived from such samples[3]. However, the replication of all sequences from independent PCRs and the uncorrected pairwise distance between the consensus sequences from the two laboratories (1.4%) suggest that few of the mutations on the DRC60 lineage are damaged-induced. Moreover, our relaxed clock methods are likely to be fairly robust to the presence of such mutations in one lineage[9]. Nevertheless, additional old sequence data would be helpful for resolving what impact, if any, this possible source of error had on the slightly earlier dates we calculated compared to previous estimates that did not include early calibration points[2,8,14,15]. Interestingly, the best-fit model for the data set that excluded ZR59 and DRC60 (**Table 1**) gave a TMRCA estimate, 1933 (1919-1945), very similar to that of Korber *et al.*[2]. This suggests that the inclusion of the old sequences, rather than the vagaries associated with a

much shorter alignment than that analyzed by Korber *et al.*, might explain the discrepancy. It is also worth noting in this context that one earlier study, using sequences from the DRC only[16], produced dating and demography estimates very similar to ours. Overall, there is broad agreement between all these studies in spite of differences in data and methods.

Our estimation of divergence times, with an evolutionary timescale spanning several decades, together with the extensive genetic distance between DRC60 and ZR59, indicate that these viruses evolved from a common ancestor circulating in the African population near the beginning of the 20th century; TMRCA dates later than the 1930s are strongly rejected by our statistical analyses. The topology of the HIV-1 group M phylogeny provides further support for this conclusion. Unlike ZR59, which is basal to subtype D[1], DRC60 branches off from the ancestral node of subtype A/A1 (**Figs. 2**, **S1**, and **S2**). Thus, phylogenetically distinct subtypes (and/or their progenitors) were clearly already present in the DRC by this early time point (**Fig. 2**). Notably, DRC60 and ZR59 cluster with other strains from the same geographical region, and basal to other members of their respective subtypes, a pattern consistent with the hypothesis that the subtypes spread through lineage founder effects worldwide, while a more diverse array of forms remained at the site of origin in Africa[17,18].

The reservoir of the ancestral virus still exists among wild chimpanzee communities in the same area on the African continent[19]. Humans acquired a common ancestor of the HIV-1 M group by cross-species transmission under natural circumstances[20], most likely predation[21]. The Bayesian skyline plot (**Fig. S2**), which tracks effective population size through time, suggests that HIV-1 group M experienced an extensive period of relatively slow growth in the first half of the 20th century. A similar pattern has been inferred using sequences sampled only in the DRC[16]. This pattern, and the short duration between the first presence of urban agglomerations in this area and the timing of the most recent common ancestor of HIV-1 group M (**Fig. 3**), suggests that the rise of cities may have facilitated the initial establishment and the early spread of HIV-1. Hence, the founding and growth of colonial administrative and trading centers like Kinshasa[22] may have enabled the region to become the epicenter of the HIV/AIDS pandemic[23].

The archival banks of Bouin's-fixed paraffin-embedded tissue specimens accumulated by many hospitals in west-Central Africa provide a vast source of clinical material for viral genetic analysis. As with the 1918 Spanish Influenza pandemic virus[24,25], a deep perspective on the evolutionary history of HIV-1 using sequences resurrected from the earliest cases in Africa could yield important insights into the pathogenesis, virulence, and evolution of pandemic AIDS viruses.

## Methods

### Archival samples

Each individual block carried an original paper identification number permanently embedded in the paraffin. Laboratory books listed the corresponding identification numbers sequentially, and included the patient's age, sex, department of hospitalization, tissue type, and date of sampling. Block identification number, sampling date, and tissue type were transcribed onto an Excel spreadsheet, and the blocks were indexed, transferred into plastic boxes, and photographed.

The results of the quantitative RT-PCR assay indicated that the integrity of the RNA preserved in these 27 samples ranged from moderate to undetectable, a range typical of Bouin's-fixed specimens[3]. The human RNA found in the 1960 lymph node biopsy sample that was found to be HIV-1 RNA positive was of relatively good quality. The Ct values

(quantitative PCR data available from the authors upon request) were as low or lower (better) than more recent (1980-90) paraffin-embedded tissues that have yielded short HIV-1 RNA amplicons[3].

Three formalin-fixed paraffin-embedded necropsy specimens were obtained from (i) a Canadian patient who died in 1997 (CAN97), (ii) a Congolese woman who died in Belgium in 1981 and was retrospectively identified as an AIDS patient (BE81), and (iii) a Congolese man who died in Belgium in 1985 (BE85). The latter two cases were presumably infected in Zaire (now DRC). The phylogenetic reconstruction shows that their viral sequences are most closely related to modern sequences from the Democratic Republic of the Congo, while the Canadian specimen yielded subtype B sequence, as predicted (**Figs. 1**, **2**, **S1**, and **S2**).

### RNA isolation and reverse transcription

Between 5 and 10 microtome sections, 5-10 μm in thickness, or an approximately equivalent amount of tissue shaved from each block with a disposable scalpel blade, were used for each digestion/extraction, as described[3]. Rigorous attention was given to preventing cross-contamination between samples by cleaning the outer surface of each block with a bleach solution, using fresh microtome/scalpel blades for each sectioning of each block, discarding the first few (exposed-surface) sections, and by performing the work in a room physically isolated from any human or HIV-1 PCR-product DNA. A 48-hour digestion period (24h at 65°C, 24h at 75°C) was used. Post extraction nucleic acids were eluted into 100 μL elution buffer AE and stored frozen at minus 80°C until required for analyses.

Reverse transcription was performed simultaneously for (i) the *gag, pol,* and human *B2M* RNA fragments, (ii) *env* fragments 3-10, and (iii) *env* fragments 11-14 (**Table S1**), with SuperScript III used according to the manufacturer's instructions. The protocol was as described[3] except that alternating 50°C and 55°C incubation periods of 30 minutes were used for a total of 6 hours.

### Amplification, cloning, and DNA sequencing

The cDNA was PCR amplified in 25μL reactions, using 0.1μL Platinum Taq HiFi enzyme (Invitrogen), 250μM DNTP mix, 2mM MgSO4, 1x PCR buffer, 0.4μM per primer, and 2μL cDNA for the *gag* and *pol* reactions, or 1μL for the *env* ones, with annealing temperatures of 60°C (*gag* [55 cycles]) or 55°C (*pol* [50 cycles]; *env* [60 cycles]). Full details are available from the authors upon request.

After amplification, the PCR-product DNA was visualized by agarose gel electrophoresis then purified using Zymoclean DNA Clean and Concentrator – 25 spin tubes (Zymo Research Corp, Orange, CA). PCR-product DNA was inserted into vector pCR 2.1-TOPO using the TOPO TA Cloning Kit (Invitrogen). The University of Arizona Genomic Analysis and Technology Core Facility resolved the DNA sequence of the vector inserts on an Applied Biosystem 3730xl DNA Analyzer using ABI Big Dye 3.1 chemistry (Applied Biosystems, Foster City, CA). Nearly identical protocols were followed for the independent replication of the DRC60 results at Northwestern University.

### Alignments

We downloaded the 2006 full-length HIV-1 sequence alignment from the Los Alamos National Laboratories HIV sequence database[26]. We retained only non-recombinant HIV-1 group M A-K subtype sequences (excluding G) and removed sequences suspected *a priori* of unusual evolutionary dynamics (such as those associated with the IV drug user epidemic in Eastern Europe and those with *nef*-deletions, both of which exhibit abnormally slow evolutionary rates). We also reduced the size of the subtype B and C clades, which are

heavily over-sampled relative to the others, by keeping only the first 5 sequences from any year/country then randomly removing sequences until the sample size was similar to other subtypes. This procedure left a total of 156 sequences. We then manually aligned the consensus sequence from the 12 regions amplified from DRC60, plus the 4 regions available for ZR59, to the full-length sequences. These short regions (**Fig. 1A** and **Table S1**) were then concatenated into an alignment 994 nucleotides in length (**Table 1**). The four *env* fragments from DRC60 that overlapped with available data from ZR59 were concatenated into an alignment 163 nucleotides in length. Matching alignments with DRC60 and ZR59 removed were also constructed. All the alignments are available from the authors upon request.

## MrBayes analyses

We used a general time-reversible nucleotide substitution model with gamma-distributed rate heterogeneity among sites and performed four independent runs of 20 million steps, sampling every 2000 steps. Examination of the MCMC samples with Tracer v1.4[9] indicated convergence and adequate mixing of the Markov chain with estimated sample sizes in the thousands. We discarded the first 2 million steps from each run as burn-in, and combined the resulting MCMC samples for subsequent estimation of posteriors. The 50% majority rule consensus tree (**Fig. S1**) is shown rooted on the branch identified by the rooted-tree method in BEAST v1.4.7, described below; however, the group M rooting was not relevant to any dating analysis. We also estimated phylogenies using the same data set under neighbor-joining and maximum likelihood methods and the same substitution model. The DRC60 sequences fell in the same topological position as with the BMCMC methods, with short root-to-tip genetic distances, consistent with the MrBayes results (**Fig. S1**). All data and trees available from the authors upon request.

We used the posterior tree sample to test the hypothesis that the terminal nodes of the DRC60 sequences were closer to the inferred A/A1 ancestral node by calculating the proportion of sampled trees where the A/A1 node-to-tip distances were smaller for these sequences than for the three modern sequences from the DRC in the same clade (**Fig. 1B**).

To predict the date of sampling based on the phylogenetic properties of the DRC60 sequences we also plotted the branch lengths (A/A1 node to tips) against the time of sampling for all A/A1 sequences excluding DRC60 and calculated the best fit for the linear regression of genetic divergence against the year of sampling of the viruses[27]. We calculated the mean and 95% HPD of the predicted sampling date of each DRC60 consensus sequence based on its node-to-tip distance and the inferred regression line calculated for each of 100 trees sampled by MrBayes.

## Bayesian MCMC inference of phylogeny using BEAST v1.4.7

We used the Bayesian methods described by Drummond *et al.*[9,10], which allow for the co-estimation of phylogeny and divergence times under a “relaxed” molecular clock model, as implemented in BEAST v1.4.7[9]. All analyses were performed under an uncorrelated lognormal relaxed molecular clock model, using a general time-reversible nucleotide substitution model with heterogeneity among sites modeled with a gamma distribution. We investigated each demographic model (constant population, exponential growth, expansion growth, logistic growth) as well as a Bayesian skyline plot coalescent tree prior[13], a general, non-parametric prior that enforces no particular demographic history. We employed a piecewise-linear skyline model with 10 groups. We then compared the marginal likelihoods for each model using Bayes factors estimated in Tracer v 1.4 as described[12,15]. Bayes factors represent the ratio of the marginal likelihoods of the models being compared. A large ratio

can indicate that one model is a significantly better fit to the data than another. We assessed the strength of the evidence that the best-fit model was superior to the others as described[15].

For each analysis, two independent runs of 50 million steps were performed. Examination of the MCMC samples with Tracer v1.4 indicated convergence and adequate mixing of the Markov chains, with estimated sample sizes in the 100s or 1000s. After inspection with Tracer, we discarded an appropriate number of steps from each run as burn-in, and combined the resulting MCMC tree samples for subsequent estimation of posteriors. We summarized the MCMC samples using the maximum clade credibility (MCC) topology found with TreeAnnotator v1.4.7[9] with branch length depicted in years (median of those branches that were present in at least 50% of the sampled trees) (**Fig. 2**). The Bayesian skyline plot was reconstructed using the posterior tree sample and Tracer v1.4.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Joel Wertheim and Mike Sanderson for computational assistance and Larry Jewel for providing the Canadian control specimen. The NIH/NIAID and the David and Lucile Packard Foundation funded the research.

## REFERENCES

1. Zhu TF, et al. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. Nature. 1998; 391:594–597. [PubMed: 9468138]
2. Korber B, et al. Timing the ancestor of the HIV-1 pandemic strains. Science. 2000; 288:1789–1796. [PubMed: 10846155]
3. Gilbert MTP, et al. The isolation of nucleic acids from fixed, paraffin-embedded tissues - which methods are useful when? PLoS ONE. 2007; 2:e537. [PubMed: 17579711]
4. Worobey M. Phylogenetic evidence against evolutionary stasis and natural abiotic reservoirs of influenza A virus. J. Virol. 2008; 82:3769–3774. [PubMed: 18234791]
5. Huelsenbeck JP, Ronquist F. MrBayes: Bayesian inference of phylogeny. Bioinformatics. 2001; 17:754–755. [PubMed: 11524383]
6. Worobey M. A novel approach to detecting and measuring recombination: insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. 2001; 18:1425–1434. [PubMed: 11470833]
7. Lemey P, et al. The molecular population genetics of HIV-1 group O. Genetics. 2004; 167:1059–1068. [PubMed: 15280223]
8. Gilbert MTP, et al. The emergence of HIV-1 in the Americas and beyond. PNAS. 2007; 104:18566–18570. [PubMed: 17978186]
9. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology. 2007; 7:214. [PubMed: 17996036]
10. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006; 4:e88. [PubMed: 16683862]
11. Salemi M, de Oliveira T, Ciccozzi M, Rezza G, Goodenow MM. High-resolution molecular epidemiology and evolutionary history of HIV-1 subtypes in Albania. PLoS ONE. 2008; 3:e1390. [PubMed: 18167549]
12. Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 2001; 18:1001–1013. [PubMed: 11371589]
13. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 2005; 22:1185–1192. [PubMed: 15703244]
14. Sharp PM, et al. The origins of acquired immune deficiency syndrome viruses: where and when? Phil. Trans. R. Soc. Lond. B. 2001; 356:867–876. [PubMed: 11405934]

15. Salemi M, et al. Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. FASEB J. 2001; 15:276–278. [PubMed: 11156935]
16. Yusim K, et al. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Phil. Trans. R. Soc. Lond. B. 2001; 356:855–866. [PubMed: 11405933]
17. Vidal N, et al. Unprecedented degree of human immunodeficiency virus type 1 (HIV-1) group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. J. Virol. 2000; 74:10498–104507. [PubMed: 11044094]
18. Rambaut A, Robertson DL, Pybus OG, Peeters M, Holmes EC. Human immunodeficiency virus phylogeny and the origin of HIV-1. Nature. 2001; 410:1047–1048. [PubMed: 11323659]
19. Keele BF, et al. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science. 2006; 313:523–526. [PubMed: 16728595]
20. Worobey, M. The origins and diversification of HIV. In: Volberding, PA.; Sande, MA.; Lange, J.; Greene, WC., editors. Global HIV/AIDS Medicine. Saunders Elsevier; Philadelphia: 2008. p. 13-21.
21. Hahn BH, Shaw GM, De Cock KM, Sharp PM. AIDS as a zoonosis: scientific and public health implications. Science. 2000; 287:607–614. [PubMed: 10649986]
22. Hance Population, WA. Migration, and Urbanization in Africa. Columbia Univ. Press; New York: 1970. p. 209-297.
23. Chitnis A, Rawls D, Moore J. Origin of HIV type 1 in colonial French Equatorial Africa? AIDS Res. Hum. Retrov. 2000; 16:5–8.
24. Taubenberger JK, et al. Characterization of the 1918 influenza virus polymerase genes. Nature. 2005; 437:889–893. [PubMed: 16208372]
25. Tumpey TM, et al. Characterization of the reconstructed 1918 Spanish Influenza pandemic virus. Science. 2005; 310:77–80. [PubMed: 16210530]
26. Leitner, T., et al., editors. HIV Sequence Compendium. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory; NM: 2005. (http://www.hiv.lanl.gov.)
27. Drummond A, Pybus OG, Rambaut A. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitology. 2003; 54:331–358.

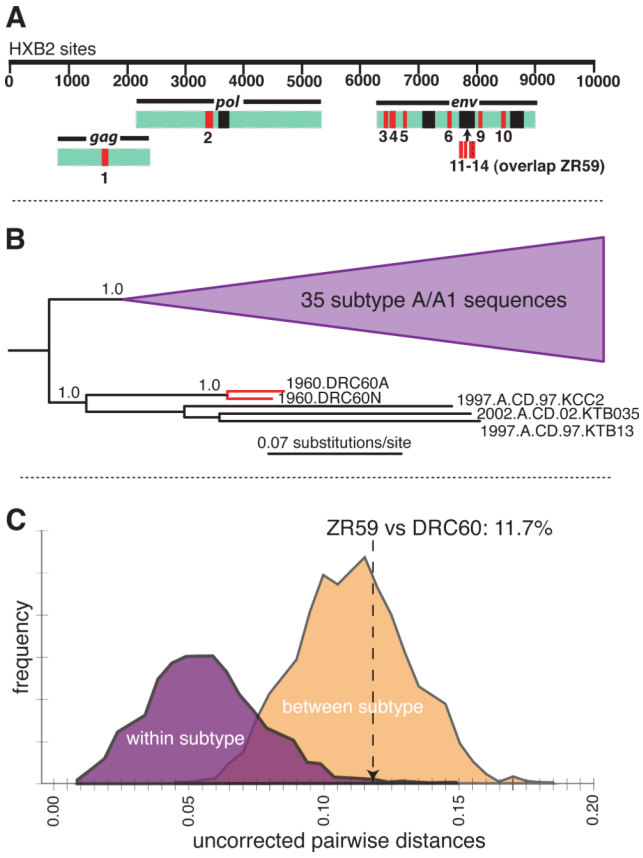


**Fig. 1.**

**(A)** The HIV-1 genome fragments that were successfully amplified from DRC60 (red) and available for ZR59 (black). The numbering for the HIV-1 sequences corresponds to the HXB2 reference sequence (**Table S1**). **(B)** The A/A1 subtree from the unconstrained (no molecular clock enforced) BMCMC phylogenetic analysis. Figure S1 depicts the complete phylogenetic tree (50% majority rule consensus tree of the posterior sample, with branch lengths averaged across the sample). Posterior probabilities are shown on nodes with support > 0.95. 1960.DRC60A is the University of Arizona consensus sequence, and 1960.DRC60N is the Northwestern University consensus sequence (i.e. the sequences independently recovered in each of the two laboratories). **(C)** Smoothed histograms of within- (A2, A/A1, B, C, D, F1, F2, H, J, K) and between-subtype distances.

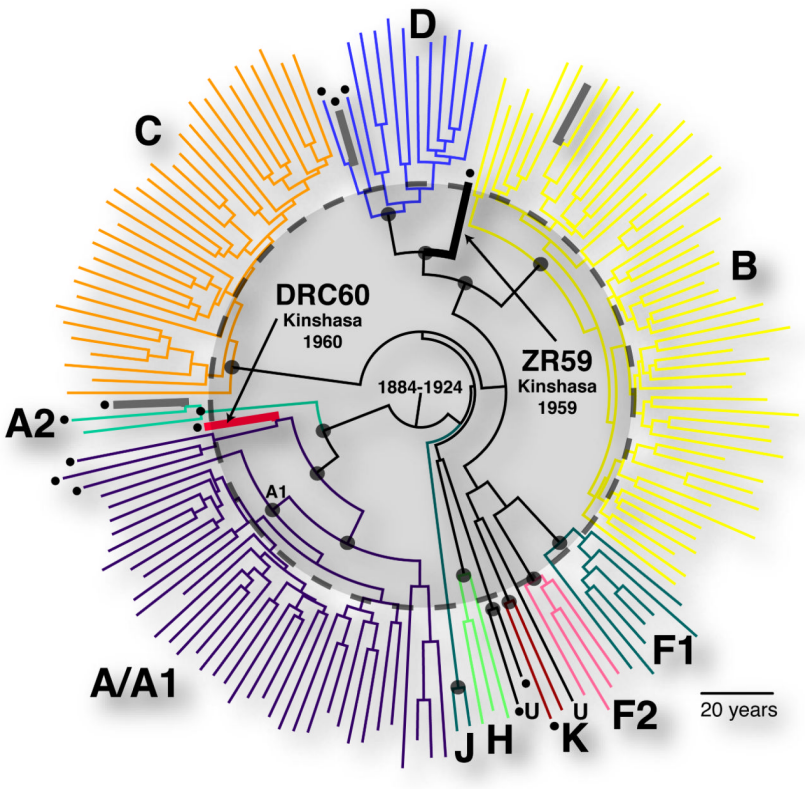


**Fig. 2.**

Maximum clade credibility topology inferred using BEAST v1.4.7 under a Bayesian skyline plot tree prior. Branch lengths are depicted in unit time (years) and represent the median of those nodes that were present in at least 50% of the sampled trees. DRC60 (red), ZR59 (black), and the three control sequences from paraffin-embedded specimens from known AIDS patients (gray) are depicted in bold. The 95% HPD of the TMRCA is indicated at the root of the tree. Nodes (sub-subtype and deeper) with posterior probability of 1.0 are marked with a large circle. Unclassifiable strains are labeled ‘U’. Sequences sampled in the DRC are highlighted with a bullet at the tip. DRC60 and the two control sequences from the DRC each form monophyletic clades with previously published sequences from the DRC, whereas the Canadian control sequence clusters, as expected, with subtype B sequences. The dashed circle and shaded area show the extensive HIV-1 diversity in Kinshasa in the 1950s. Figure S2 shows the tree in rectangular form with taxon labels.

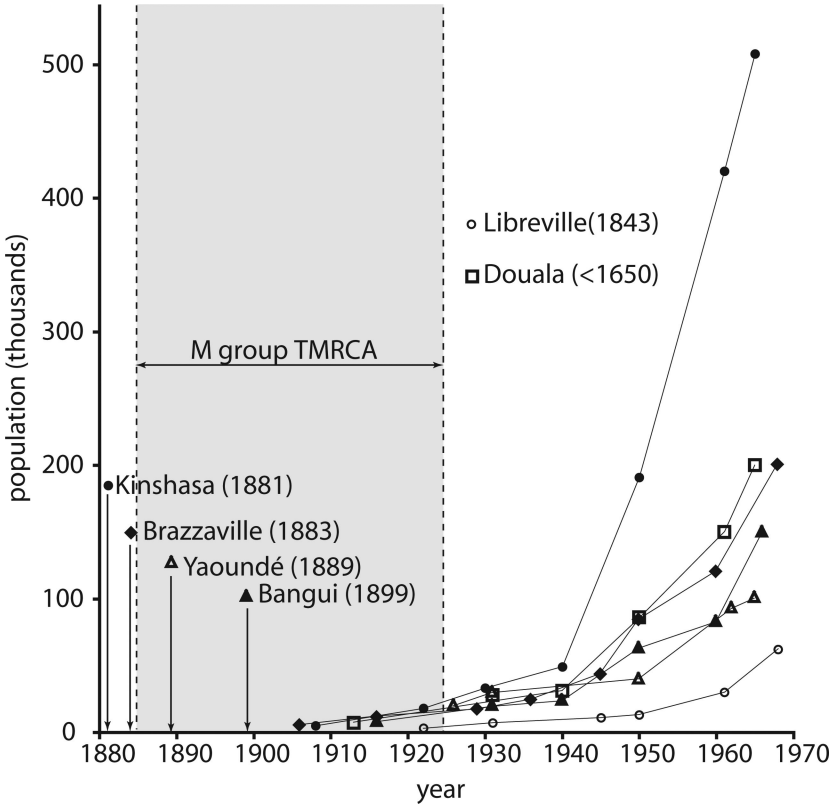


**Fig. 3.**

The origin and growth of the major settlements near the epicenter of the HIV-1 group M epidemic. In the countries surrounding the putative zone of cross-species transmission[19] (current-day Cameroon, Central African Republic, DRC, Republic of Congo, Gabon, and Equatorial Guinea) there was not a single site with a population exceeding 10,000 until after 1910. The founding date of each major city in the region is listed beside its name. Most were founded only shortly before the estimated TMRCA of group M. The demographic data come from reference 23.

## Table 1

HIV-1 M group TMRCA estimates from BEAST analyses under different coalescent tree priors.

| Coalescent tree prior | DRC60 & ZR59 excluded[a] | DRC60 & ZR59 included |
|---|---|---|
| ***Constant*** | **1933 (1919-1945)**[b] [0.0][c] | **1921**[d] **(1908-1933)** [0.0] |
| ***Exponential*** | 1907 (1874-1932) [-3.5 ±0.8] | 1914 (1891-1930) [-2.1 ±1.5] |
| ***Expansion*** | 1882 (1834-1917) [-2.7 ±0.8] | **1902 (1873-1922)**[d] [-1.6 ±1.5] |
| ***Logistic*** | 1913 (1880-1937) [-2.3 ±0.8] | 1913 (1891-1930) [-3.2 ±1.5] |
| ***BSP***[e] | 1882 (1831-1916) [-2.7 ±0.8] | **1908 (1884-1924)**[d] [-0.4 ±1.5] |

[a] Concatenated *gag-pol-env* fragments available for either or both of ZR59 and DRC60 (994 nucleotides total, 507 from DRC60).

[b] Median and 95% highest probability distribution of TMRCA.

[c] log10 Bayes factors difference in estimated marginal likelihood (± estimated standard error) compared to the coalescent model with strongest support.

[d] TMRCAs for the best-fit model and models not significantly worse than it are written in bold text.

[e] Bayesian skyline plot tree prior.